Rev: November 13, 2012

DRAFT

Chapter from a book, *The Material Theory of Induction*, now in preparation.

Simplicity as a Surrogate[1]

John D. Norton
Department of History and Philosophy of Science
Center for Philosophy of Science
University of Pittsburgh
http://www.pitt.edu/~jdnorton

1. Introduction

The idea is found almost everywhere, from the most prosaic to the most abstruse settings. Choosing the simpler option speeds you to the truth. In ordinary life, when the lights go out, we choose the simpler hypothesis that the electrical power has failed. We discard the more complicated hypothesis that all the bulbs malfunctioned at the same moment and, worse, that each malfunctioned for a different reason. In cosmology, we choose the simpler hypothesis that the same physical laws obtain here as in distant places and epochs, even though we cannot rule out that they may differ in parts quite remote from us.

Do these judgments implement a universal principle of inductive inference that says:

> If two hypotheses are each adequate to the phenomena, the simpler is more likely true.

My goal in this chapter is to deny the efficacy of any such universal principle of inductive inference. For the material theory of induction entails that no such rules are efficacious. To explain the popularity of appeals to simplicity, I will urge that good invocations of simplicity are really veiled references to background facts or assumptions whose content functions to license

[1] My thanks to Fellows in the Center for Philosophy of Science, Fall 2012, for discussion of an earlier draft of this chapter.

the relevant inductive inference. The apparently singular appeal to simplicity actually masks an appeal to such a diversity of context dependent facts that no univocal meaning can be attached to it.

This is the sense in which simplicity is a surrogate. In so far as it is epistemically efficacious, the short and snappy invocation of simplicity is really a surrogate for background facts or assumptions. These background facts do the real epistemic work and, commonly, are much harder to capture in a comparably short slogan. There will be cases in which these backgrounds resemble one another so that a common idea of simplicity appears to be invoked. However the extent of these cases will always be limited. As we move farther afield, we will encounter cases in which the backgrounds differ sufficiently for the similarity to fail. In general, there is no well-specifiable, universally applicable, epistemically efficacious principle of simplicity in inductive inference.

This analysis is a deflationary analysis of simplicity that runs counter to the celebration of simplicity in the scientific literature. It does have a small pedigree in the philosophical literature. It is the view of simplicity long defended by Elliott Sober. His Sober (1988) uses emphasized text to summarize his view as:

> Whenever a scientist appeals to parsimony to justify the conclusion that one hypothesis is more reasonable than another in the light of observational data, substantive assumptions about the world must be involved. In practice, parsimony cannot be "purely methodological." (p.40)

and then more compactly:

> Appeal to simplicity is a surrogate for stating an empirical background theory. (p.64)

The following section provides a brief illustration of how apparently epistemically efficacious invocations of simplicity are really indirect appeals to background facts. Section 3 brackets off two cases of lesser interest in which simplicity offers only pragmatic gains. They are the cases in which simplicity is urged as an efficient search heuristic and in which simplicity is demanded merely to give a compact summary of past experiences.

The two sections that follow develop and deflate two primary senses of simplicity. The first principle, discussed in Section 4, expresses simplicity in a count of entities or causes. The classic statement is Ockham's razor: "Entities must not be multiplied beyond necessity." It fails

as a principle of parsimony, I will argue, since there is no clear way to count the numbers of things to be minimized. The principle is reinterpreted as truism of evidence, that one should not infer to more entities than the evidence warrants, where this evidential warrant is understood materially. The second principle of parsimony, discussed in Section 5., requires us to infer to hypotheses whose description is simple. This principle fails as an independent principle since modes of description vary. These variations greatly affect the descriptive simplicity of hypotheses. This form of the principle can only guide us if we fix the mode of description and the guidance will be good only if that mode is properly adapted to the prevailing facts.

Section 6 will examine in more detail the most popular illustration in the philosophical literature of the use of simplicity, curve fitting. The invocation of simplicity in standard curve fitting, I argue, is a surrogate for specific background facts. They are: the obtaining of a particular model of how error in data confounds some true curve; that the parameterization used is suitably adapted to the backgrounds facts; and that, in the strongest cases of this adaptation, the hierarchy of functional forms fitted corresponds to background assumptions on the presence, likelihood and strength of certain processes. Ascending the hierarchy is not authorized by some abstract principle that tells us to proceed from the simpler to the more complex. Rather it is a successive accommodation of the curves fitted to the most likely or strongest processes and then to those less so. The concluding two sections 7 and 8 illustrate this last adaptation of the hierarchy in the examples of fitting orbits to observed positions in astronomy and the harmonic analysis of tides.

2. How it Works: The Birds

Just how can simplicity serve as a surrogate for background facts? Here is an easy illustration. Imagine that you are walking on the beach over sand washed smooth by the ocean waves. As you walk over a clear expanse of smooth sand, you notice a track left by a bird (Figure 1).

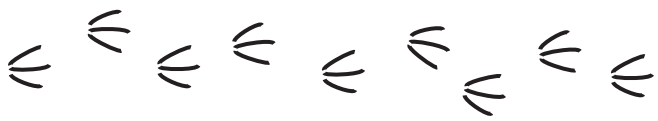


Figure 1. Bird Tracks

The prints are clear and unmistakable. You can just see how the bird waddled over the sand—left, right, left, right—leaving the prints. But why assume that it was just one bird? Perhaps the left foot prints were made by a one-legged bird that hopped awkwardly over the sand. Then a second one-legged bird, this time having only the right leg, pursued it, leaving the right footprints in just the right place to simulate the waddle of a single two-legged bird. Or perhaps there was a large flock of one-legged birds, each of which touched down on the sand just once, all perfectly coordinated to leave the track.

Each hypothesis explains the track. However we do not take the various, one-legged bird hypotheses seriously. How might we defend this judgment? The one bird hypothesis is by far the simplest. In ordinary discourse, merely declaring that might be a sufficient defense. If our methodology is at issue, then merely declaring that it is the simplest might not be enough to secure it. If we need more, we can turn to the great Isaac Newton. At the start of Book III of his magisterial *Principia* he asserted four "Rules of Reasoning in Philosophy," which would guide the subsequent analysis. The first two rules are (Newton, 1726, p.398):

> Rule I
>
> *We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.*
>
> To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.
>
> Rule II
>
> *Therefore to the same natural effects we must, as far as possible, assign the same causes.*
>
> As to respiration in a man and in a beast; the descent of stones in *Europe* and in *America*; the light of our culinary fire and of the sun; the reflection of light in the earth, and in the planets.

These two rules remain the clearest and firmest pronouncement of a methodological principle of parsimony in science.

Applied to the birds, Rule I tells us immediately that we should use the one bird hypothesis, for it is a truth that there are two-legged birds and their behavior is sufficient to

explain the tracks. We do not need the many bird hypothesis, so it should not be admitted. In so doing, we conform with Rule II by assigning the same cause, a single bird, to the many footprints.

So far, all is well. Simplicity has provided the principled justification for our intuitive judgment. That will not last, however. We now proceed farther down the beach and come to a place where the smoothed sand is criss-crossed by very many tracks, in evident confusion (Figure 2).

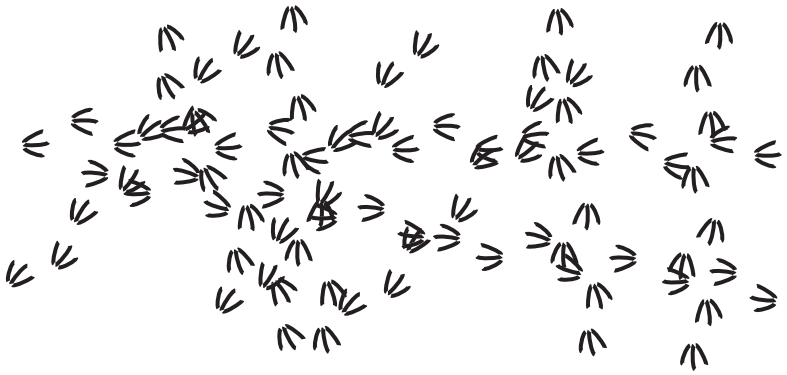


Figure 2. More Bird Tracks

We would normally posit that many birds alighted on the sand, each leaving just one track. However there is another hypothesis: the tracks were left by just one, very busy bird. It waddled over the sand; flew to another spot; waddled some more; and so on, until the final set of tracks was formed.

A mechanical application of Newton's rules leads us directly to the one busy bird hypothesis. We are, as before, assigning the same cause, one bird, to the same effects, the one set of tracks. Few of us would accept this outcome. We would be satisfied with one bird hypothesis for the single track but expect a good analysis to return a many bird hypothesis for this case of many tracks. We would surely reason something like this. In the case of the single track, we rule out the many, one-legged bird hypothesis because we know that one-legged birds are rare and, if they were on the beach, it is unlikely that they would follow each other around in just the way needed to produce a single track. For the case of many tracks, we know that it is possible for one bird to be very busy and produce the multiple tracks. However we rarely if ever see such a lone, busy bird, whereas flocks of birds producing tracks like this are quite common.

These further reflections show that our initial analysis was not merely based on the invocation of simplicity. We chose the one bird hypothesis for the single track on the basis of our

relatively extensive knowledge of birds. It is a shared knowledge, so we generally feel no need to explain in tedious detail why we rule out other possible, but unlikely hypotheses: two one-legged birds hopping, many one-legged birds alighting once, a mutant four-legged bird, and so on. We can dismiss all these far-fetched notions with a breezy wave towards the simplest hypothesis.

In short, what we offer as a conclusion governed by some general principle of parsimony is really a conclusion dictated by our knowledge of background facts. We use an appeal to simplicity as a convenient way of circumventing the need to explain in detail these background facts, whose details can become quite convoluted. My claim here is that all epistemically efficacious invocations of simplicity have this character.

3. Pragmatic Justifications of Simplicity

Let us return to the standard view that treats a preference for simplicity as a methodological principle of universal character. What justifies it? What precisely does the principle assert? My interest in simplicity is restricted to the case in which simplicity functions epistemically as a marker of truth; we are to choose the simpler hypothesis or theory because, we are assured, it is more likely to be true. I will argue below that a principle of this form has no precise content and no good justification. However, before we proceed, we need to dispense with two distracting special cases that lie outside my concerns. In them, simplicity is sought merely for pragmatic reasons.

3.1 Simplicity for Economy of Search

In seeking to understand some new phenomenon, scientists commonly deal with many hypotheses or theories. How should they go about searching among them and testing them? A common recommendation is that they should start with the simplest hypothesis or theory. The simplest are easiest to deal with and, if they are incorrect, likely to be refuted by new evidence sooner than a more complicated one.

In the 1920s, it was found that distant galaxies recede from us with a speed that increases with distance. In 1929, Hubble proposed that the speed of recession was linearly proportional to the distance. In principle, he could have fitted a complicated, tenth order polynomial function to his data. The linear dependency, however, was easier to deal with formally. If it is the wrong

relation, new data would be likely to show the error much faster than with a more complicated function. A tenth order polynomial is able to contort itself to fit a larger range of data, so that considerably more data may be needed to refute it.

This sort of circumstance is common. One of the simplest hypotheses concerning an ailment is that it is caused by a specific pathogen. Famously, in the mid nineteenth century, John Snow was able to localize the cause of a cholera outbreak in London to drinking tainted water, drawn from a public water pump at Broad Street. More recently, the cause of AIDS—acquired immune deficiency syndrome—has been identified in the HIV virus. Once the simple hypothesis was pursued, it was readily affirmed. Were definite pathogens not responsible, the simple hypothesis could likely have been ruled out fairly quickly by the appearance of cases in which no exposure to the conjectured pathogen was possible. Matters are quite different with ailments such as cancer. Multiple factors can make a cancer more likely, including carcinogenic chemicals, ionizing radiation, certain viruses and even specific genes. Dealing with this multiplicity of causal factors and discerning which are operating when, is considerably more difficult.

These simple observations have been incorporated into analyses of scientific discovery. Karl Popper (1968, Ch. VII) urged that science proceeds through a continuing cycle of the conjecture of new hypotheses and their refutation. He identified the simpler hypotheses with the more falsifiable. It follows that the cycle advances faster if the scientists investigate more falsifiable hypotheses, that is, simpler hypotheses. A mathematically more sophisticated analysis of the role of simplicity in heuristic search has been provided by Kelly (2007). In the context of a formal learning theoretic analysis of the evidence-guided search for hypotheses, he shows that favoring simpler hypotheses is a more efficient way of getting to the truth.

How are these considerations relevant to our present concerns? One might ground the favoring of simplicity in searching in two related suppositions: that nature is ontically simple or that nature is descriptively simple in our languages. In both these cases, further discussion must be deferred to later sections of this chapter, where I argue that both suppositions are epistemically efficacious only in so far as they make indirect appeals to background assumption.

However these assumptions are not needed to ground the heuristic recommendation. It is still good advice to investigate the simplest first in a world that is indifferent to the simplicity of hypotheses. Whether the world is more likely to give us a linear function or a tenth order polynomial, the linear function will still be dealt with more easily and more quickly. Whether

ailments are more likely to be caused by a single pathogen or by many factors, we still proceed most expeditiously by checking the single pathogen hypothesis first.

In short, simplicity can remain a good heuristic in hypothesis searching without any need for nature to be governed by a general principle of simplicity or parsimony.

3.2 Simplicity as Mere Economy of Expression

Ernst Mach famously held the view the scientific laws were merely compact summaries of our experience. He said in an 1882 address to the Imperial Academy of Sciences in Vienna "The goal which it [the intellect] has set itself is the *simplest* and *most economical* abstract expression of facts." (Mach, 1898, p. 207) The idea can be put crudely as follows. Galileo asserts that the distance fallen by a body varies with the square of time of fall. In Mach's view, all that Galileo is allowed to assert is that each pair of distances and times we have measured for falling bodies conforms to this relation.

In so far as this is all that is asserted, then the role of simplicity is merely that of convenience. One seeks the least troublesome way of summarizing the facts at hand. In more modern terms, the exercise is essentially one of data compression. We could report all the numerical data pertaining to the fall of many bodies; or we could report merely that these data all conform to Galileo's relation without loss of anything that matters.

This may seem an extreme view that is, nowadays, well out of the philosophical mainstream. However much engineering practice conforms to it. That is because engineering commonly deals with systems dependent on many variables and the systems are sufficiently complicated that a fundamental analysis is precluded. To deal with this problem, the behavior of the system is measured experimentally under widely varying circumstances and the collected data reduced to as compact a form as possible.

One of the best-known examples is the treatment of fluid flow in pipes. Even this simple problem involves many variables: the fluid's speed, density and viscosity; the pressure drop in the paper; and the pipe's diameter and surface roughness. Once the flow becomes turbulent, this empirical approach is the only tractable one. Moody (1944) presented a now famous chart summarizing the outcomes of many experiments. See Figure 3.

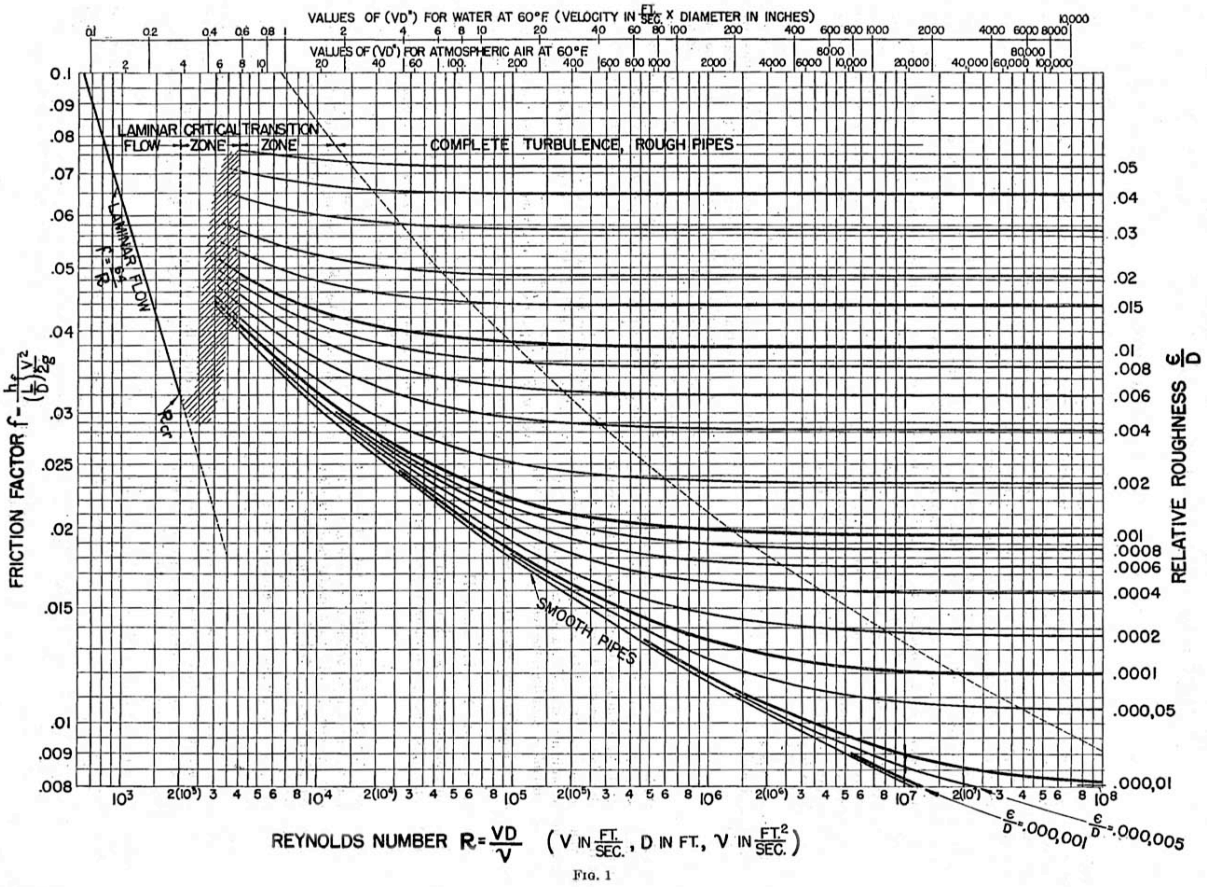
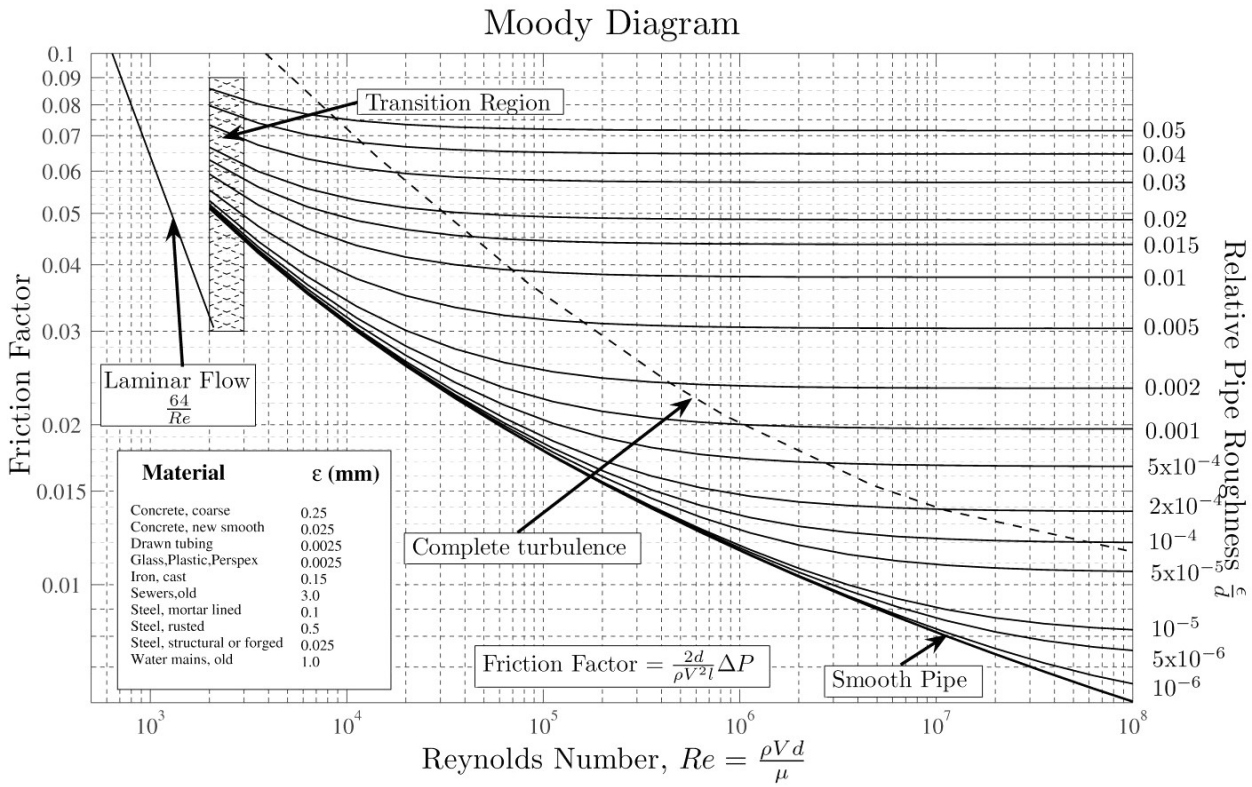


Fig. 1

Figure 3. Moody Chart

(Note: This Moody diagram is the original from the 1944 paper. It may still be under copyright. If so, it can be replaced by



from

http://en.wikipedia.org/wiki/File:Moody_diagram.jpg

which is released under the GNU Free Documentation License.)

In this one chart, one can read the pressure drop associated with the flow of fluid of specified speed, density and viscosity in a pipe of specified diameter and surface roughness.

In so far as the chart merely summarizes the outcomes of experiments already performed, it is free of entanglement with the concerns of this chapter. One need make no reference to background facts when one reports that a simple formula generates a curve that happens to pass through the data points near enough. I will complain shortly of the ambiguity in the notion of simplicity. That ambiguity is untroubling here. We can use any formula that generates a curve that fits the data well enough. The choice is purely pragmatic.

This purely pragmatic use of simplicity is an extreme case. I believe that it is rarely and possibly never realized in all purity. The examples above do not realize it fully. The Moody chart is a summary of past experience. But it is also a great deal more. Its primary use is as an instrument of prediction. The presumption is that, if an engineer constructs a piping system with flow rates, fluid densities, and so on, matching conditions covered by the chart, then the relevant

curve will reflect the pressure drop that will be found. That can only happen if the chart is properly adapted to broader facts about fluid flow in pipes in general.

These facts have the character of simplicity assumptions. We must assume that the variables included are all that matters. Temperature does not enter into the chart explicitly; it is assumed that thermal effects are fully captured by the unrepresented dependence of density and viscosity on temperature. We must assume that the curves fitted to the data points interpolate correctly between them so that the chart makes good predictions for cases whose precise combination of variables have never been observed.

In so far as the descriptions seek to go beyond past experience, they seek the type of epistemic success to which the subsequent discussion applies.

4. Principles of Parsimony: Ontic Simplicity

The notion that parsimony can successfully guide us epistemically has many expressions and one might despair of categorizing them all successfully. There is, however, a broad division between ontic simplicity and descriptive simplicity. I will discuss ontic simplicity first and later turn to descriptive simplicity.

In this ontic version of the principle, we are guided to the truth by favoring accounts that posit the fewest entities or processes in the world. The locus classicus of this notion is "Ockham's razor." Its now universal formulation is

Entia non sunt multiplicanda praeter necessitatem.

Entities must not be multiplied beyond necessity.

Curiously this formulation is not to be found in the writings of the fourteenth century scholastic, William of Ockham. His closely related pronouncements include[2]

It is useless to do with more what can be done with fewer.

A plurality should not be assumed without necessity.

It has been an historical puzzle to locate the source of the popular formulation.[3] Another puzzle is that Ockham's name should be so exclusively attached to this maxim of simplicity, for it was

---

[2] Quoted from Maurer (1999), p. 121. The Latin is "Frustra fit per plura quod potest fieri per pauciora." "Pluralitas non est ponenda sine necessitate."

an idea that, according to Mauer (1999, p. 121), was used commonly from the thirteenth century, after being gleaned from Aristotle.

The greater puzzle is why modern thinkers would look to a fourteenth century scholastic for this sort of guide. His understanding of the demand of parsimony was rather different from its modern use in science. He felt it not binding on god. He felt, as Maurer (1999, p. 120) reports, that "God is not bound by it; he does many things by more means which he could do by fewer, and yet this is not done uselessly, because it is God's will."[4]

The better-formulated statement of the sentiments in Ockham's razor are Newton's two rules of reasoning as quoted in the last Section. The notion is advanced explicitly as a rule of reasoning; and Newton provides a justification. "Nature does nothing in vain." and "Nature is pleased with simplicity, and affects not the pomp of superfluous causes." The justification is dressed in anthropomorphic garb. Nature, surely, is never literally pleased or displeased. Without that garb, Newton is enjoining us to infer to the simpler case of fewer causes because the world is simpler, harboring fewer causes. This is a factual claim about the world.

This justification seems routine. We find it reported in Aquinas (1945, p. 129) (writing before Ockham's birth):

> If a thing can be done adequately by means of one, it is superfluous to do it by means of several; for we observe that nature does not employ two instruments where one suffices.

Its Difficulties

This ontic version of the principle of parsimony faces many difficulties. The most immediate is that we have no general prescription for how to count the entities, processes or

---

[3] See Thorburn (1918).

[4] My reaction to this puzzle is that we have fallen into introducing a defective principle of parsimony with the faux dignity of a pedigreed Latin maxim in the hope that it will deflect a skeptical demand for justification. We may be unprepared to justify just why the two entity hypothesis is better than the three. But we can plant the idea that it is what *everyone* thinks and has done so since the fourteenth century as a way of stalling skeptical challenges. On a superficial survey of its use, it appears that this subterfuge is working pretty well.

causes to which the principle is applied. It is not hard to find ambiguity sufficiently severe as to compromise the principle.

How do we count entities when we compare a continuum and a molecular theory of gases? The continuum theory represents the gas as a single, continuous fluid. The molecular theory represents it as a collection of very many molecules, of the order of 10^{24} in number for ordinary samples of gases. Do we count one entity for the continuum gas and 10^{24} for the molecular gas, so that the molecular gas posits many more entities? Or do we invert the count? A continuum is indefinitely divisible into infinitely many parts.[5] The molecular gas consists of finitely many molecular parts. Has the continuum now infinitely many more parts than the molecular gas?

This discussion in terms of entities can be converted into the causal conception of Newton's rules. What causes the pressure exerted by a gas jet impinging on a surface? Do we count the impact of the continuum gas as one cause? Or do we count an infinity of causes for the infinitely many impacts of its infinitely many parts?

What of the justification for this ontic principle? Newton asserts the world is simpler in employing fewer causes. That assertion is empty in so far as the counting of causes is ill-defined. However even setting that concern aside, the claim is still unsustainable. Nature is not simple. Traditional alchemical theories posited three or four elements in an attempt to account for chemical appearances. We now know that this count is far too low. A tractable chemistry requires over ninety elements. Perhaps Nature is pleased with chemistry, but surely not for the simplicity of the count of elements.

The existence of isotopes is especially damaging to Newton's justification. For one can explain the chemistry of carbon quite well just by assuming that there is one element, carbon. Hence, Newton's rules urge, we should infer to there being just one element carbon since Nature "affects not the pomp of superfluous causes." That is, we should infer to the one element and not to the existence of multiple types of chemically identical carbon, because that is the way Nature

---

[5] In statistical physics, this gives a continuous entity such as a field infinitely many degrees of freedom and is responsible for the "ultraviolet catastrophe" of classical electromagnetic fields. Correspondingly, a molecular gas has finitely many degrees of freedom as a result of its finitely many molecules.

is. Yet that is not the way Nature is. Carbon exists in multiple, chemically identical isotopes, Carbon-12, Carbon-13 and Carbon-14. The recommendation to infer to just one type of carbon may well be good advice as far as the chemistry of carbon is concerned. I do not wish to impugn this recommendation or to suggest that an inductive rule is defective because it sometimes leads us astray. That is part of the risk we take whenever we carry out inductive inference. Rather the issue here is the justification. While the recommendation may be good, it cannot be justified by a supposition that factually there is just one type of carbon. Factually, there is not.

Rescue by the Material Theory: the Principle as an Evidential Truism

This ontic form of the principle of parsimony is troubled. Yet it has figured and continues to figure prominently in successful science. There must be something right about it. The clue to what is right lies in the ambiguous qualifications found in every formulation. We are not to multiply entities "beyond necessity." We are to admit no more causes than are "sufficient to explain…" We assign the same cause "as far as possible." These qualifications mean the principle is not self-contained. Something more supplies the sense of necessity, possibility and sufficiency.

Newton's formulation gives us the clearest indication of what that something is. We are not to proceed beyond that which is "sufficient to explain the[ir] appearances." That gives us some flexibility. We can add to the causes we admit as long as they are sufficient to explain the appearances. We are not to go beyond that sufficiency. Since we routinely infer inductively to that which we count as explaining the appearances, this amounts to telling us to infer to no more than that for which we have inductive authorization. Understood this way, the principle is revealed to be a truism of inductive inference, which says:

> We should not infer to more than that for which we have good evidence.

It is a corollary of another truism: we should infer inductively only to that for which we have good evidence.

How did an inductive truism become enmeshed with the muddle of the metaphysics of simplicity? The key relevant fact is that the truism is not an independent inductive principle; it is a meta-inductive principle. That is, it is not a principle *of* an inductive logic. Rather, it is a principle *about* how other inductive logics should be used. That this is so is harder to see if one conceives of inductive inference formally. The principle is entangled with assertions about how

the world is factually. If one understands inductive inference materially, however, that entanglement is expected. Moreover it clarifies how the original principle can be a good epistemic guide.

We can see this entanglement in Newton's first use of his Rules I and II. In Book III of *Principia*, Proposition IV Theorem IV asserts that the force of gravity that draws objects near the earth's surface is the same force that holds the moon in its orbit. He assumes that the force acting on the moon intensifies with decreasing orbital radius according to an inverse square law, as it does with other celestial objects. It follows that were the moon to be just above the earth's surface, it would fall to earth with the same motion as ordinary bodies fall by gravity. He continued:

> And therefore the force by which the Moon is retained in its orbit becomes, at the very surface of the Earth, equal to the force of gravity which we observe in heavy bodies there. And therefore (by Rule 1 & 2) the force by which the Moon is retained in its orbit is that very same force which we commonly call gravity; for were gravity another force different from that, then bodies descending to the Earth with the joint impulse of both forces would fall with a double velocity…

Newton here invokes his rules to complete the inference. However the inference is already fully controlled by a factual presumption: that the matter of the moon is the same as the matter of the earth and, if brought to the surface of the earth, would behave like ordinary terrestrial matter.

That factual assumption already authorizes Newton's conclusion and he gives the reason. Were there to be some additional celestial force that acts on the matter of the moon but not on ordinary terrestrial matter, then the moon would fall with double the motion of ordinary terrestrial matter. That contradicts the assumption that the matter of the moon behaves just like that of the earth. This is a kind of simplicity assumption: contrary to ancient tradition, there is no difference between terrestrial and celestial matter. But its comprehension and use make no appeal to abstract metaphysical notions of simplicity. It is a specific factual statement and it powers the inductive inference, as the material theory requires.

We also see this entanglement in the illustrations Newton supplies for his Rules. To illustrate Rule II, he considers the "light of our culinary fire and of the sun." We are to assign the same cause to both. We now know that this is an erroneous conclusion. Culinary fires generate light from combustion; the sun generates light by a different process, nuclear fusion. What makes

the inference appear unproblematic for Newton is that he is really relying on a tacit background assumption: that it is very unlikely that there is a process that produces intense light other than combustion. That fact powers the inductive inference.

In short, this ontic formulation of the principle of parsimony fails as a universal principle of inductive inference. It is too vague to be applied univocally and efforts to give it a foundation in a supposed general, factual simplicity of the world founder. Its successes, however, can be understood in so far as it is the meta-inductive principle that one should infer inductively to no more that for which one has good evidence. The assertions of simplicity are veiled invocations of relevant facts that authorize the inductive inference, in accord with the material theory.

5. Principles of Parsimony: Descriptive Simplicity

These descriptive versions of the principle do not directly address the numerical simplicity of entities or causes. Instead we are enjoined to favor the simplest *description* of processes pertaining to them. It may not be possible to effect an absolute separation between the descriptive and the ontic versions of the principles of parsimony. For descriptive simplicity is tacitly supposed to reflect some sort of ontic simplicity. However the explicit focus on language introduces sufficient complications to need a separate analysis.

The best-known application of descriptive simplicity is curve fitting. In its simplest form, we are given a set of data points, that is, many measured values of a variable x and a variable y. These are represented as points on a graph, as shown below. We then seek the curve that fits them best. It is routine in curve fitting to start with a constant relation, then a linear one, then a quadratic, and so on, seeing how much better is the fit of the curve as we proceed to higher order polynomials, as shown in Figure 4.

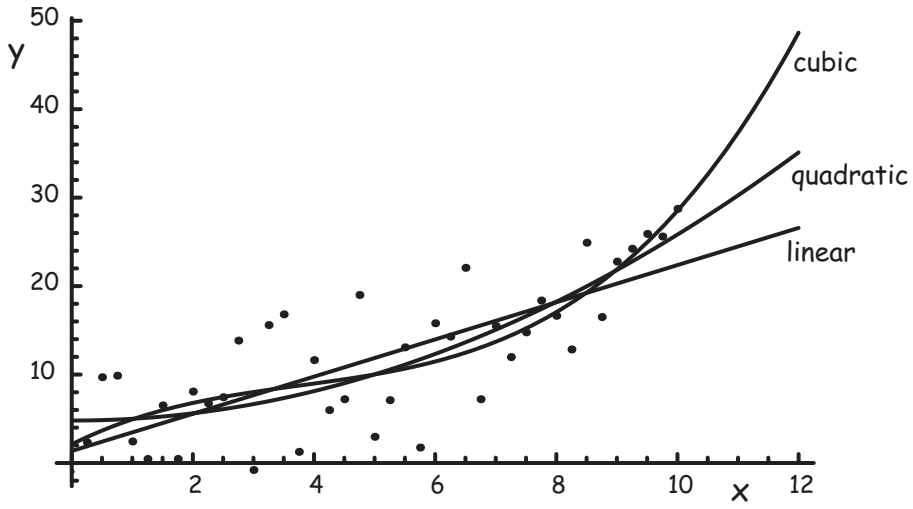


Figure 4. Polynomial Curve Fitting

The fit will always improve as we increase the order. The higher the order of the polynomial, the larger the repertoire of curves available and hence the more likely we are to come close to the data points.

Eventually, however, this greater flexibility will cause trouble. For the data is routinely assumed to be a compound of the true curve sought and confounding error. If the true law sought is merely a linear curve, the error will scatter the data around the true straight line. Higher order polynomial curves will have little trouble adapting to the random deviations due to noise. This will lead the fitted curve to deviate from the true curve as it responds to the vagaries of the noise. Figure 5 shows the best fit of linear and eighth order polynomial curves to a data set.

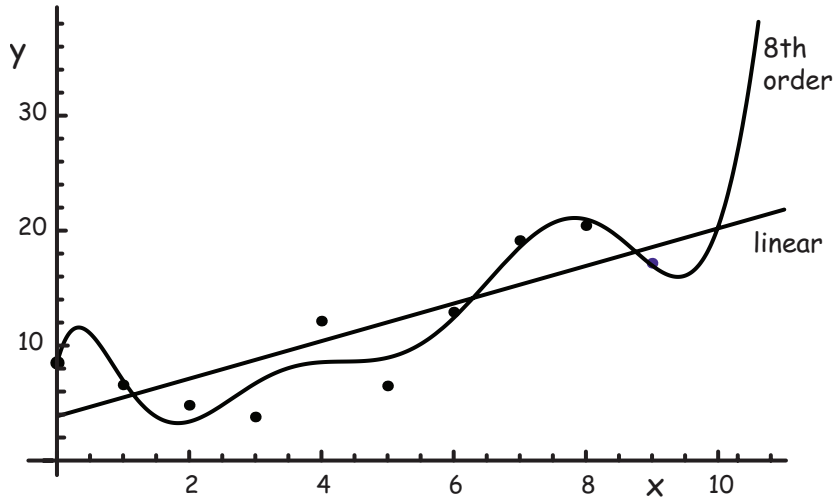


Figure 5. Overfitting

The apparently better fit of the higher order curves is spurious. This phenomenon is known as "overfitting."

The primary burden in curve fitting is to find a balance of the two effects: the simplicity of the curves that fit less well against the better fit of more complicated curves that overfit. The simplicity of a curve is derived from its description. The polynomial family consists of smaller, nestled families of curves:

Constant: $y = a$

Linear: $y= a + bx$

Quadratic: $y = a + bx + cx^2$

Cubic: $y = a + bx + cx^2 + dx^3$

Quartic: $y = a + bx + cx^2 + dx^3 + ex^4$

and so on.

That is, the formulae that describe the curves have more parameters as we proceed to the less simple, higher order polynomials. The constant curve has one parameter, a. The linear curve has two parameters a and b. The quadratic curve has three parameters, a, b and c. And so on. Built into the practice of curve fitting is a simplicity postulate: favor those curves whose descriptions require fewer parameters.

The preference for simpler descriptions has been applied more broadly. Perhaps its best-credentialed proponent is Albert Einstein. The laws of fundamental theories of physics employ constants. Newton's theory of gravitation employed the gravitational constant G. Einstein's special theory of relativity employed the speed of light c and his general theory employs both c and G. Quantum theory employs Planck's constant h, as well as numerous quantities characteristic of the various particle interactions, such as the charge of the electron e. The standard model of particle physics now employs nearly 20 such constants. Some of these constants can be absorbed into the system of units used. The speed of light c can be suppressed merely by measuring distance in light years; then the speed of light reduces to unity.

Einstein (1949, p. 61-63) grounded his hope for a physics free of all further constants in a belief in the simplicity of nature

> If one considers this [suppression] done, then only "dimension-less" constants could occur in the basic equations of physics. Concerning such, I would like to state a theorem which at present cannot be based upon anything more than upon a faith in

> the simplicity, i.e., intelligibility, of nature: there are no *arbitrary* constants of this kind; that is to say, nature is so constituted that it is possible logically to lay down such strongly determined laws that within these laws only rationally completely determined constants occur (not constants, therefore, whose numerical value could be changed without destroying the theory).

While the freedom from these constants reflects something factual in the structure of the world, Einstein expresses it in terms of the descriptions of that structure, that is, in terms of the constants appearing in the equations that describe it. Just as curve fitting should favor smaller numbers of parameters, Einstein favors laws with the fewest arbitrary parameters.

These sentiments come from Einstein later in his life. By then he had abandoned his earlier allegiance to positivistic approaches. He had become a mathematical Platonist and that was a doctrine, he assured us, he had learned from his experiences in physics.[6] His 1933 Herbert Spenser lecture, "On the Methods of Theoretical Physics," offers an explicit and powerful manifesto:

> Our experience hitherto justifies us in believing that nature is the realisation of the simplest conceivable mathematical ideas. I am convinced that we can discover by means of purely mathematical constructions the concepts and the laws connecting them with each other, which furnish the key to the understanding of natural phenomena. Experience may suggest the appropriate mathematical concepts, but they most certainly cannot be deduced from it. Experience remains, of course, the sole criterion of the physical utility of a mathematical construction. But the creative principle resides in mathematics. In a certain sense, therefore, I hold it true that pure thought can grasp reality, as the ancients dreamed.

Einstein continues to detail how we can use mathematical constructions to make discoveries in physics:

> The physical world is represented as a four-dimensional continuum. If I assume a Riemannian metric in it and ask what are the simplest laws which such a metric can satisfy, I arrive at the relativistic theory of gravitation in empty space. If in that

---

[6] For an account of precisely how Einstein's experience with general relativity led him to this, see Norton (2000).

> space I assume a vector-field or an anti-symmetrical tensor-field which can be derived from it, and ask what are the simplest laws which such a field can satisfy, I arrive at Maxwell's equations for empty space.

The recipe is one of descriptive simplicity. In each context one writes the simplest admissible equations and thereby recovers the law.[7]

Its Difficulties

The difficulty with any principle expressed in terms of descriptions is that it can be rendered incoherent by merely altering the descriptive system used. In so far as the simplicity principle of curve fitting merely requires us to favor the curves with fewer parameters, it is unsustainable. Merely rescaling the variables used can overturn its judgments completely, as will be illustrated in the following section.

The idea that we get closer to the truth by writing mathematically simpler laws has the imprimatur of Einstein. However it unsustainable for the same reasons that trouble curve fitting. Judgments of just what is descriptively simple are too malleable. Einstein's own general theory of relativity illustrates the problem. When the theory first came to public notice after the eclipse tests of 1919, it was notorious for its abstruse difficulty. The eminent astronomer George Ellery Hale confided in correspondence:[8]

> …I confess that the complications of the theory of relativity are altogether too much for my comprehension. If I were a good mathematician I might have some hope of forming a feeble conception of the principle, but as it is I fear it will always remain beyond my grasp.

---

[7] Here are the technical details for those who want them. The simplest non-trivial structure in the derivatives of the metric tensor g_{ik} is the Riemann curvature tensor, R^{i}_{kmn}. Its vanishing requires the flatness of spacetime, which is too restrictive. The vanishing of its unique first contraction, R_{ik}, is the Einstein gravitational field equation for empty space. The vector field is the vector potential A_i and the tensor field mentioned is the Maxwell field tensor $A_{i;k} - A_{k;i}$. Setting its divergence to zero returns the source-free Maxwell equations.

[8] February 9, 1920. Quoted in Clark (1984, pp. 299-300).

The *New York Times* of November 19, 1919, reported an incredible tale, reflected in the partial headline "A book for 12 wise men":

> When he [Einstein] offered his last important work to the publishers he warned them there were not more that twelve persons in the whole world who would understand it, but the publishers took it anyway.

The fable took root. It is repeated in the publisher's introduction to Lorentz's (1920, p.5) popularizations of relativity theory.

As the decades passed, general relativity was absorbed into mainstream physics and opinions began to migrate. By the 1970s, the standard textbook for the theory came to a startlingly different conclusion (Misner, Thorne and Wheeler, 1973, pp. 302-303):

> "Nature likes theories that are simple when stated in coordinate-free, geometric language."…According to this principle, Nature must love general relativity, and it must hate Newtonian theory. Of all theories ever conceived by physicists, general relativity has the simplest, most elegant geometric foundation … By contrast, what diabolically clever physicist would ever foist on man a theory with such a complicated geometric foundation as Newtonian theory?

How is a reversal of this magnitude possible? The key is the centrality of "coordinate-free, geometric language." One finds general relativity to be the simpler theory when one adopts the appropriate language. As a result, the principle of descriptive simplicity, as enunciated by Einstein, is incomplete. Without a specification of the right language to be used, it can give no direction at all.

Perhaps one might hope that somehow mathematics provides the natural descriptive language for our science and physics in particular. A cursory glance at the interplay of mathematics and science shows things to be different. There is no unique language for physical theories. New physical theories commonly appear mathematically difficult and even messy. That is followed by efforts by mathematicians and the scientists themselves to simplify the presentation and manipulations of the theory. As I have argued in Norton (2000, pp. 166-68), what results are new mathematical methods and new formulations of the theories that become successively simpler.

Newton's development of his mechanics employed the ancient methods of Euclidean geometry. His contemporaries required considerable insight and facility in geometry to follow

and emulate his difficult demonstrations and proofs. Over the subsequent centuries, Newton's theory was re-expressed in terms of algebraic symbols and the calculus. Many of what were once abstruse results became natural and simple. Quantum mechanics developed in the first quarter of the 20th century. The theory that resulted in the late 1920s was a complicated mess of differing approaches and techniques: matrix mechanics, wave mechanics, Dirac's c- and q-numbers. Subsequent efforts showed all the theories to be variant forms of a single theory that found its canonical mathematical formulation in von Neumann's 1932 classic, *Mathematical Foundations of Quantum Mechanics*. Even Einstein's general relativity benefited from this reforming. His original methods did not include the now key notion of parallel displacement. This notion was introduced in response to the completion of the theory by the mathematician Levi-Civita in 1917.

Rescue by the Material Theory: Adaptation of Language as a Factual Judgment

As before, we must acknowledge that there is something right in the idea that descriptive simplicity is a guide to the truth. What is right is already clear from the discussion above. The real epistemic work is done in finding and developing a language or descriptive apparatus appropriate to the systems under investigations. What makes that apparatus appropriate is precisely that the truths concerning the system find simple expression in it. Then it is automatic that seeking simple assertions in the language or descriptive apparatus leads us to the truth.

That is, the principle of descriptive simplicity guides us to the truth in so far as the language we use is properly adapted to the background facts. Hence what is really guiding us is not some abstract notion of simplicity but merely the background facts as reflected in our choice of descriptive language. This inductive guidance from background facts is, of course, precisely what is called for by the material theory of induction. This idea will be illustrated with the example of curve fitting in the next section.

6. Curve Fitting and the Material Theory of Induction

As a mode of discovery, curve fitting is based on the idea that fitting a simple curve to data can guide us to the truth. The material theory of induction requires that these inductive inferences are warranted by background facts. Here I will describe in greater detail the character of these background facts. We will see that the vague and incomplete injunctions to seek the

simpler curve translate into more precise constraints, expressed in terms of these background facts. The characteristics of background facts found in many but not all cases of curve fitting can be grouped under three headings, as below.

The Error Model

When curve fitting is used to seek some underlying truth or law, the presumption is that the data to which the curve is fitted have been generated by a standard error model of the form:

Error laden data = true curve + error noise

That curve fitting operates with data produced by this model is so familiar that it is easy to overlook its importance. The techniques of curve fitting are designed to strip away confounding noise. Thus the assumption of the standard error model must be true if these techniques are to guide us towards the truth.

A quick way to see its importance is to consider the curve fitting problem shown in Figure 6. We seek the value of the quantity a that gives the best fit of y = 1/[log(x)-a] to the data.

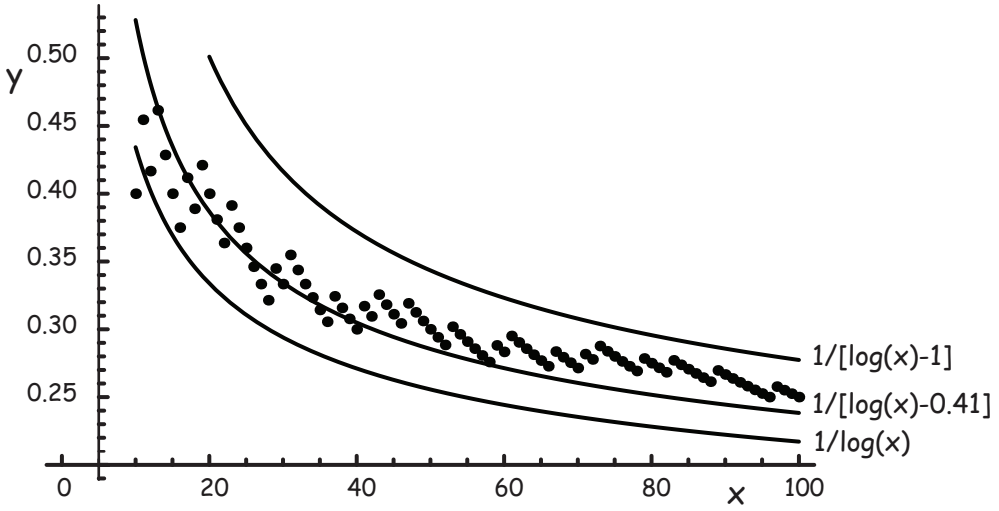


Figure 6. A Curve Fitting Problem

The optimum value turns out to be the value shown, a = 0.41.

Superficially, the problem looks traditional. However this curve fitting problem does not conform to the standard error model. The data represent the density of prime numbers between x= 10 and x=100. The first datum at x=10 reports the number of primes between 1 and 10. There are four of them: 2, 3, 5 and 7, so that y= 4/10 = 0.4. The prime number theorem assures us that the density y(x) approaches 1/log(x) for large x. A corollary is that the density also approaches

1/[log(x) – a] for *a* some constant, for the two quantities, 1/log(x) and 1/[log(x) – a] approach one another for large x. The curve fitting problem is to ascertain which value of a gives the best fitting curve for values of x in the range 10 to 100 covered by the data. The result is 0.41 as shown.[9]

Instead of the standard error model, this problem conforms to a non-standard error model in which truth and error are permuted:

True data = error laden curve + error noise

This means that, epistemically, the exercise is different. We are not seeking truth. We already have the complete truth in the data that report the true density of prime numbers. Instead we are seeking a summary that has least deviation from the truth, where the notion of "least deviation" is one we can choose conventionally. In this case, I chose a fit that minimized the sum of squared deviations.

Engineering applications, such as the Moody diagram above, illustrate a second way that we may deviate from the standard error model. In so far as we are merely seeking a compact summary of past experience, there is no real error model in use at all, for there is no hidden truth sought. Our ambitions, however, are rarely so limited. For example, as noted above, the Moody diagram is typically not intended merely as a compact historical report. It is also intended as a predictive tool. In so far as that is the case, the standard error model is presumed.

However the practice is somewhat protected from the full rigors of the model by the fact that engineering practice rarely requires perfectly exact prediction of pressure drops in pipes. A prediction correct to within a few percent is more than sufficient for most applications. This affords great protection when we seek predictions for conditions that interpolate between those used to create the original chart. Fitting just about any family of curves will interpolate to the requisite level of accuracy. In effect we are conforming the data to a weaker model:

Error laden data = near enough to true curve + error noise

Near enough to true is good enough for interpolation.

This protection is lost when we seek to extrapolate to new conditions outside those used to create the diagram. For then two curves that each interpolate among the data equally well may diverge markedly when we extend beyond the condition in which the data were collected. Then

---

[9] This is specifically for primes in the range specified. The optimum value for all primes is a = 1.

we need to find which is the true curve on pain of uncontrolled errors entering our predictions. This divergence is illustrated in Figure 4 above. The polynomials in the figure interpolate the data comparably well in the range of x = 0 to x = 10. They diverge rapidly outside the range in which the data was collected, giving markedly different results in the range x = 10 to x = 12.

The Parameterization

Descriptive simplicity can only be a good epistemic guide to the truth, I have urged, if the language of description is chosen so that the truths correspond to simple assertions. In the case of curve fitting, that condition translates into a matching with background facts of the parameterization used and the family of its functions from which the curves are drawn.

We are free to rescale the quantities used to describe our measurements; and we do. We may compare cars by their speeds or, equivalently, by the times they take to cover some fixed distance. Since one parameter is the inverse of the other, fitting some family of curves to speed will in general give different results from fitting the same family of curves to times. This reparameterization is common. In acoustics, we measure loudness of sounds in decibels, which is a logarithm of the power of the sound. In astronomy we measure the apparent magnitude of the stars on a scale that is the logarithm of the intensity (that is, the energy flow per unit area).

To see how the parameterization we choose makes a difference, we will develop an example in which the data are generated by a true linear relation y = x, as in Figure 7. The data has been simulated with very little noise, so the best fitting straight line[10] comes so close to y=x.

[10] The best fitting straight line is y = -0.000403379 + 0.996917 x. It is not shown in Figure 7 since it is too close to the true curve y=x to be separated visually.

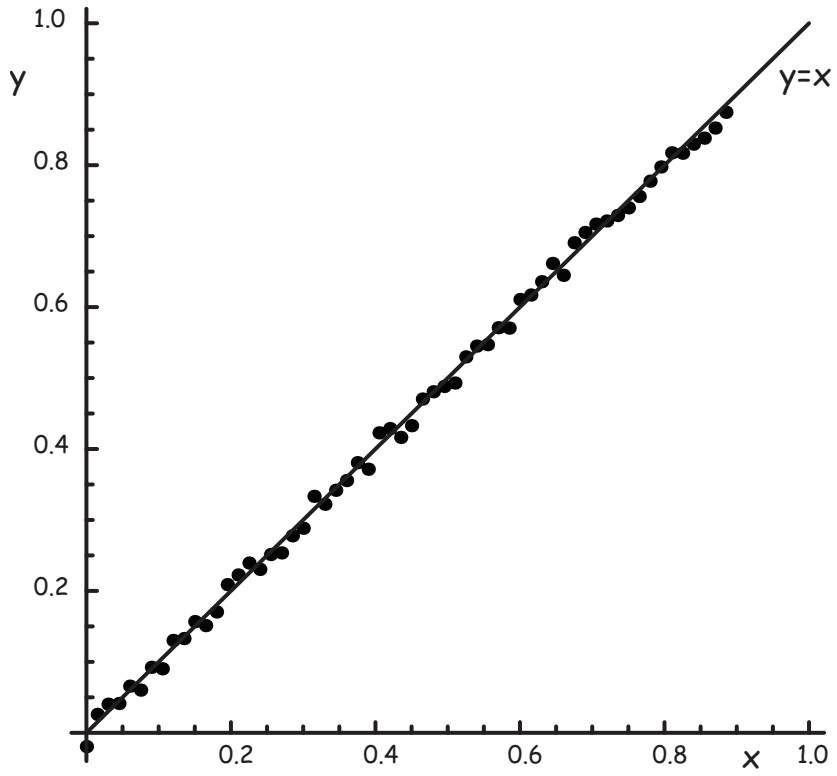


Figure 7. Data generated from true curve y=x

Rescaling the x variable to $z = \sin^{-1}(x)$, means that the true curve rescaled to $y = \sin(z)$. However a polynomial curve fit between y and z will never return this curve, for y=x is equivalent to a polynomial of infinite order in z:

$$y = \sin(z) = z - (1/3!)z^3 + (1/5!)z^5 - (1/7!)z^7 + \ldots$$

A curve fitting algorithm that proceeds up the family of polynomials in z will necessarily halt at some finite order and so cannot return the true curve. Finding polynomials of best bit for the rescaled data of Figure 7 shows how poorly the polynomial fit performs. The best fitting linear, quadratic, cubic and quartic polynomial curves interpolate the data well. However, as Figure 8 shows, they fail immediately on extrapolation beyond the domain x = 0 to x = 0.9 in which the data was generated.[11]

[11] The domain x= 0 to x = 0.9 corresponds to z = 0 to $z = \sin^{-1}(0.9) = 1.12$. The largest x shown, x=1, corresponds to $z = \sin^{-1}(1) = \pi/2 = 1.57$.

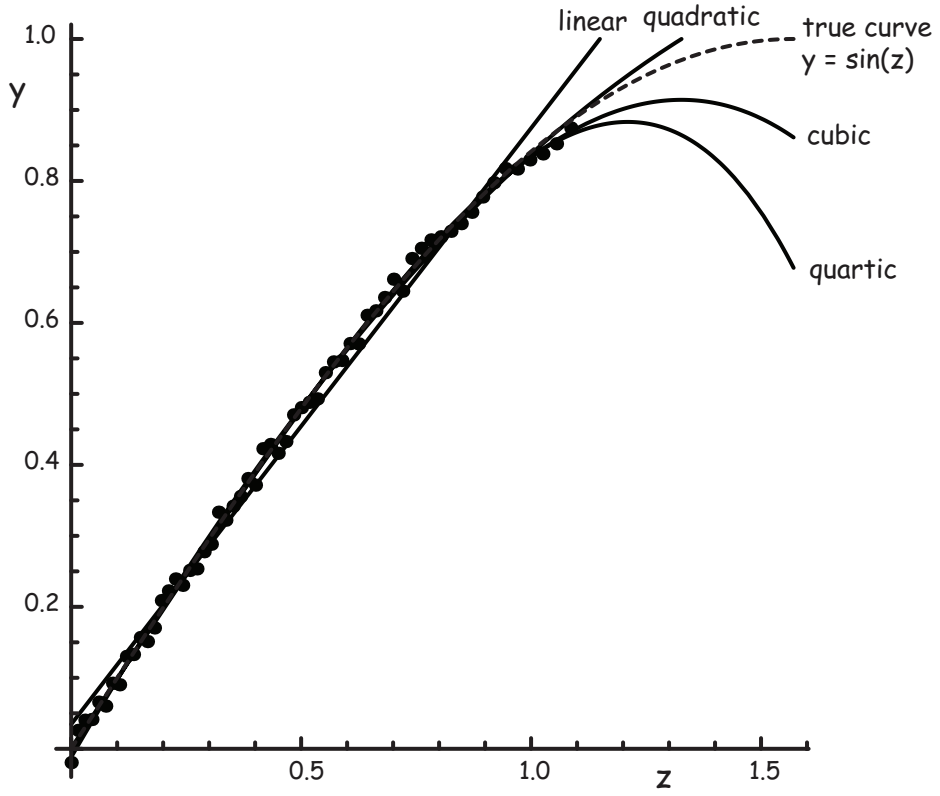


Figure 8. Failure of polynomial curve fit on reparameterized data

The problem is the same in the reverse direction. If y=z is the true curve, then this true curve corresponds to an infinite polynomial if we parameterize the data using x, for

$$y = \sin^{-1}(x) = x + (1/6)x^3 + (3/40)x^5 + (5/112)x^7 + \ldots$$

for $|x|<1$. Once again ascending the family of finite polynomials will never return the true curve.

Choosing the right parameterization and family of curves amounts to properly adapting them to the background facts. If the attenuation of the intensity I(r) of some signal with distance r is due to an absorptive medium, then the signal attenuates as $I(r) = I(0) \exp(-\lambda r)$, for λ some constant. The exponential dependence of I(r) amounts to another infinite order polynomial in r. If we rescale, the relation reduces to a simple linear dependence of the logarithm of I(r) on r, for then the attenuation follows $\log I(r) = \text{constant} - \lambda r$. If, however, the attenuation is due to spreading it space, signal intensity will attenuate according to $I(r) = A/r^2$, for some constant A. This corresponds to $\log I(r) = A - 2 \log(r)$, which once again corresponds to an infinite order polynomial of in r. However, if we use both log I(r) and log(r) as our parameters, then the true curve is linear and its slope, -2, conveys the fact that the attenuation follows an inverse square law.

Perhaps the clearest example of this adaptation of the parameters and curves to the background facts arises when we have processes that are periodic in time. We should then use the time t as the parameter. The family of curves to be fitted should not be polynomials, since they are not periodic. Rather we should use the family of periodic trigonometric functions, sin(t+a), sin(2t+b), sin(3t+c), etc, where the a, b, c, … are constant phase factors. We learn from Fourier analysis, that this family is sufficiently rich to represent all periodic curves of likely interest to curve fitters.

The Order Hierarchy

We must have an adaptation of the descriptive language to background facts if descriptive simplicity is to be an effective guide to truth. In important cases the adaptation can be especially tight. Curve fitting proceeds with some collections of families of curves, such as the constant, linear, quadratic, etc. In these important cases, the families of curves fitted correspond directly to particular processes. Then fitting a curve from a family farther up the hierarchy corresponds to the inclusion of more processes in the account developed of the phenomena. Further the adaptation has to be such that curves fitted earlier in the procedure correspond to stronger or more probable processes.

This adaptation will be illustrated in the following two sections with the cases of fitting trajectories to celestial objects and the harmonic analysis of the tides.

7. Fitting Orbital Trajectories

The standard method of curve fitting is to find that curve that minimizes the sum of the squares of deviations of the curve from the data. This least squares technique was introduced around the start of the 19th century in astronomy to assist the fitting of orbits to celestial objects in our solar system. This application illustrates a tight adaptation on the curve fitting method to the background assumptions that are the surrogates of simplicity. The family of curves fitted reflects the particular trajectories that background assumptions select. Moreover ascending the order hierarchy reflects pursuit of trajectories according to their likelihood and the strength of the processes that form them.

Ellipses, Parabolas and Hyperbolas

A new celestial object—a new planet or comet, for example—is sighted. The astronomers' first task is to find the orbit that fits the positions seen. Astronomers do not follow the curve fitters' generic procedure of seeking first to fit a straight line and then proceeding up through higher order polynomials. Rather the family of curves chosen is provided by gravitation theory. The initial model is provided by the "one body problem": the motion of a free body attracted to a central point by a force that varies inversely with the square of distance r to the point. That is, the attracting force is k/r^2 for k a suitable constant. The familiar result, given early in any text on celestial mechanics,[12] is that the trajectory is one of the three conics sections: an ellipse, a parabola or a hyperbola.

Select polar coordinates (r,θ) in the plane of the orbit with the origin r=0 at the center of force in the sun and set $\theta=0$ at the perihelion, the point of closest approach to the sun in the orbit. A single formula that covers all three curves is

$$r = \frac{L}{1 + e\cos(\theta)}$$

where e is the eccentricity and L is the semi-latus rectum that, loosely speaking, fixes the width of the figure. (More precisely, it is the distance from a focus to the curve along a line perpendicular to the major axis.) We pass among the conic sections with equal semi-latus recta by changing the eccentricity e. A circle is e=0, an ellipse is $0<e<1$; a parabola is e=1; and a hyperbola is $e>1$.[13]

Figure 9 shows trajectories of the three types of conic section with equal semi-latus rectum:

---

[12] I happen to be using Sterne (1960, §1.3) here.

[13] The semi-latus rectum is related to the semi-major a by $L = a(1-e^2)$ for both an ellipse and an hyperbola if we adopt the convention that a is positive for an ellipse and negative for an hyperbola.

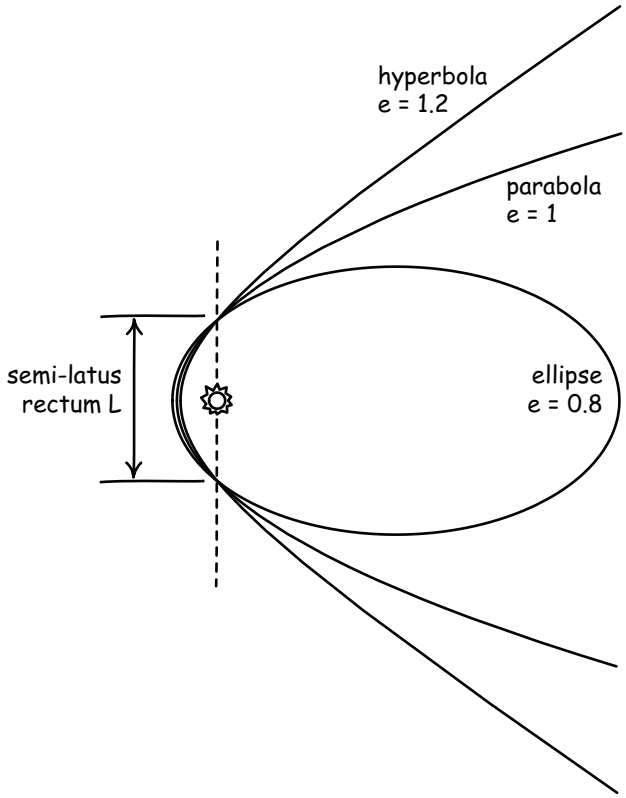


Figure 9. Conic Sections with Equal Semi-Latus Rectum

Once the semi-latus rectum L and eccentricity e are fixed, so is the orbital motion. Kepler's area law affirms that the radial line from the center of force to the object sweeps area at a constant rate with time t. That is the areal velocity $h=r^2 d\theta/dt$ is constant. This areal velocity is related to L by $h = \sqrt{kL}$. Hence the angular speed of the object in its orbit, $d\theta/dt$, is fixed once the trajectory is fixed.

Consider once again the default of the straight line of generic curve fitting. This familiar default is precluded by the background assumptions of gravitation theory. It can arise for objects in the solar system if they are moving at speeds very much greater than those normally encountered. Then the object will follow a hyperbolic trajectory with a very large eccentricity that is practically indistinguishable from a straight line.

The three conic sections provide the basis for the family of curves employed. When we allow for perturbations produced by gravitational attractions from other objects in the solar system, the family is enlarged by allowing a slight motion in the curve. For example, the major axis of the ellipse along which the object moves may rotate slowly. Representing that slow

rotation introduces further parameters and provides the full family of curves used in virtually all accounts of orbital motion.

Fitting an orbit to a celestial object involves moving up this hierarchy of curves until a suitably close fit is obtained. One might try to describe this ascent as guided by some principle of parsimony that requires starting with the simplest curve and then moving up to more complicated ones. However it is hard to see just which abstract notion of simplicity might here lead us to identify conic sections as the simplest case with straight lines an extreme case never implemented. Fortunately no such notions of simplicity are needed to explicate the procedures used. The selection of curves and their order are guided by background assumptions on the likelihood of certain trajectories and on the likelihood and strength of processes that modify them in prescribed ways.

Comets

An easy illustration is provided by the methods used to fit orbits to newly discovered comets. That is, the illustration is easy if we limit ourselves to the methods routinely used in the nineteenth century. Watson (1861, pp. 163-169) describes the methods then customary. They depend essentially on the following background fact: Comets typically have very eccentric orbits and we get to observe them when they are in the vicinity of the sun. There, as Figure 9 above suggests, it becomes quite difficult to separate the ellipses and hyperbolas with eccentricity close to unity from each other and from a parabola.[14] This fact leads to the procedure described by Watson (1861, p.164):

> It is therefore customary among astronomers, when a comet has made its appearance unpredicted, to compute its orbit at first on the supposition that it is a parabola; and then, by computing its place in advance, find from a comparison of the actual observations, whether this hypothesis is the correct one. Should it be

---

[14] The details: If the trajectory is a parabola with semi-latus rectum L, then the distance to the sun at perihelion is L/2. For a very eccentric ellipse or hyperbola, e is close to 1; that is $e = 1-\varepsilon$, where ε is small. (It is positive for an ellipse and negative for an hyperbola.) Hence $1-e^2 = 1-(1-2\varepsilon + \varepsilon^2) \approx 2\varepsilon$ to first order. Hence the semi-latus rectum $L = a(1-e^2) \approx 2a\varepsilon$. The distance to the sun at perihelion is $a(1-e) = a\varepsilon \approx L/2$, which agrees with the parabolic case.

> found to be impossible to represent the observed positions of the comet by a parabola, an ellipse is next computed and when this also fails, recourse is had to the hyperbola, which, provided the previous computations are correct in every particular, will not fail to represent the observations within the limits of their probable errors.

That is, the known fact of the high eccentricity of comets directs a choice of a parabola to fit the initial data. The astronomers then collect more data and move to a nearby ellipse and then hyperbola if the deviations from the parabola are sufficient.

The sequence of shifts reflects a definite physical assumption about the energy of the comet. Adopting an ellipse amounts to assuming that the comet's total energy—kinetic plus potential—is negative so that it is bound to the sun and can never escape. Watson does not mention it here, but I believe the decision to try an ellipse next rather than an hyperbola reflects the prevalence of comets bound in elliptical orbits. Adopting the hyperbola amounts to assuming a positive energy sufficient to enable the comet to escape the sun's gravity. The case of the parabola is the intermediate case of zero energy, which is the minimum level at which escape from the sun's gravity is possible.

There is a small element of arbitrariness in the procedure. Instead of fitting a parabola initially, the astronomers could have chosen an ellipse with an eccentricity imperceptibly different from unity. (That trajectory would be near indistinguishable from a parabola in the vicinity of the sun.) Whichever is chosen as the first curve, the choice is driven by the background physical assumption that the comet is just on the energetic edge of being gravitationally bound permanently to the sun. Further data then directs a decision to one or other side, within or beyond the edge. The selection of the curves fitted reflects this background physics.

Perturbed Orbits and New Planets

The familiar slogan is that planets orbit the sun in elliptical orbits. This slogan is almost exactly correct. A planet will trace out an almost perfectly elliptical orbit over the shorter term. If it is tracked carefully over the longer term, however, slight deviations will appear. They are sufficiently small that they can be represented as changes in the elliptical orbit that the planet is following. If, for example, the axis of the ellipse rotates in the plane of the orbital motion of a

planet, then the orbit actually traced out takes on the flower petal shape seen in Figure 10. It is an advance of the planet's perihelion, its point of closest approach to the sun, with each orbit of the sun. Or at least this is the shape traced out if we depict an advance unrealistically fast for any planet.

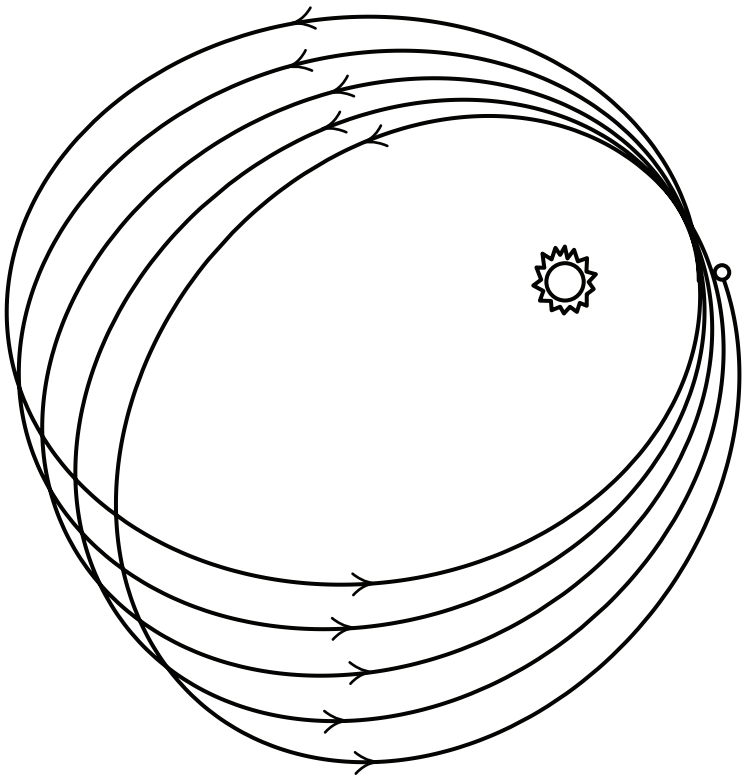


Figure 10. Advancing Perihelion Motion of a Planet

In general, the small deviations from the perfect elliptical orbits are represented by slow changes in the six Keplerian elements that characterize each elliptical orbit. The first two elements are the most familiar: the semi-major axis and eccentricity specify the shape of the ellipse. The remaining elements locate that ellipse in space and place the planet on it at the appropriate time.[15] The families of curves associated with these perturbed elliptical orbits are the ones fitted to the observed positions of the planets.[16]

---

[15] The inclination and longitude of the ascending node locate the orientation of the orbital plane holding the ellipse in relation to the reference plane, which is usually the ecliptic. The argument of the periapsis locates the orientation of the orbit's major axis in the orbital plane. Finally, the

These perturbed ellipses are fitted to the planets initially because they are found to fit. However, from the earliest time of Newtonian gravitation theory, the challenge has been to locate the physical cause for the perturbation, which is almost invariably sought in the perturbing gravitational influence of bodies other than the sun. Jupiter, the most massive of the planets, is a common source of perturbations. It exerts a large perturbing influence on Mercury for example. The axis of Mercury's orbit advances and the ellipse is more eccentric when Jupiter is in line with this axis. The axis regresses and is less eccentric when Jupiter is perpendicular to the axis.[17]

This need to give a physical foundation for the perturbed ellipses fitted is uncompromised. One might initially find that some perturbed ellipse fits the motion. However that fit remains tentative until the physical basis is located. Only then can the astronomers know how well the perturbed ellipse will continue to fit the planet's motion. More importantly, the perturbations to the ellipse can be adjusted according to the subsequent movements of the perturbing body.

Perhaps the most vivid illustration of the need for a physical basis for the changing elements of a planet's ellipse arises in the celebrated discovery of the planet Neptune. By the early 19th century, the orbit and perturbations of the planet Uranus had been established. However not all of the perturbations could be explained by the gravitational action of known planets. In 1845, Adams and Le Verrier independently pursued the possibility of another hitherto unknown planet outside the orbit of Uranus that would be responsible for the perturbations. They predicted the position of this planet. After an easy telescopic search in 1846, the planet was found and was eventually given the name Neptune.

That astronomers *require* the variant curve forms to have a physical foundation is seen most clearly when these efforts fail. The orbit of Mercury was also well established in the nineteenth century and the bulk of its perturbations could be accounted for by the gravitational effects of the other planets. However they could not be accounted for completely. Recalling the

---

mean anomaly at epoch fixes the position of the planet in its orbit at one time. (If that position is known, the dynamics of gravitation theory can be used to determine its position at other times.)

[16] Our moon's motion is greatly perturbed, so that this approach is less successful for it and other methods are used in the historically troublesome lunar theory. See Brown (1896, p. 66).

[17] Or so Airy (1884, p.113) reports.

success with Neptune, Leverrier (1859) proposed that these further perturbations could be accounted for by another new planet orbiting closer to the sun than Mercury. The new planet, which had come to be known as Vulcan, was never found.

That failure was discouraging. Nonetheless, astronomers could not abandon the idea that the perturbations were generated by some attractive mass somewhere. If the mass was not concentrated in a planet, perhaps it was distributed. Seeliger in 1906 proposed that the requisite attraction might come from two superposed oblate homogeneous ellipsoids of matter around the sun, which, it was supposed, would also explain the phenomenon of the Zodiacal light. (Jeffreys, 1916)

Another physical basis was possible. In 1895, the puzzling anomaly in Mercury's motion was reported by Newcomb (1895, p. 814) in his synoptic survey of the elements of the four inner planets. He noted an excess motion in the perihelia of the planets that was most pronounced for Mercury. Rather than leave it as a purely descriptive matter, he incorporated a physical assumption into the analysis: the gravitational attraction of the sun dilutes with distance r not as the inverse square $1/r^2$, but as $1/r^{2.00000016120}$. What resulted was the requisite correction to Mercury's perihelion motion of an advance of 43.37 seconds of arc per century and related smaller advances in the perihelia of Venus, Earth and Mars.

We might well wonder if Newcomb took this adjustment to the inverse square law seriously as a new physical hypothesis, for there is no discussion of what appears to be a quite artificial breach in an immutable law of nature. Perhaps it was merely introduced in the spirit of Osiander's preface to Copernicus' work. That is, it serves only as a useful fiction for computing the motions of the planets.

Whatever his attitude may have been, it soon turned out that Newcomb would have been correct to seek the source of the anomalous motion of Mercury in some adjustment to Newton's law. For the decisive resolution came with just such an adjustment, though hardly one Newcomb could have anticipated. In November 1915, an exhausted Einstein (1915) was putting the finishing touches onto his nascent general theory of relativity. He discovered to his jubilation that the new theory predicted precisely the anomalous advance of the perihelion of Mercury. He computed an advance of 43 seconds of arc per century, noting that the astronomers' values lay in the range of 40-50 seconds. The achievement was startling and has remained a favored and perhaps over-used illustration of how evidence can bear on theory. For Einstein recovered the

result without the need for any ad hoc hypotheses, such as a fine-tuning of the exponent in the force law for gravity.

8. Harmonic Analysis of Tides

On an oceanic coast, the level of the seawater rises and falls periodically, with about two high tides and two low tides each day. Beyond this crude description are many variations. There is some variability in the timing of highs and lows; and there is considerable variation in just how high is a high and just how low is a low tide. This variability is also somewhat periodic over longer time scales, but the exact cycles are hard to pin down precisely merely by casual inspection of some portion of a tide's history.

Accurate tidal prediction is important and even essential for coastal navigation. Since the ebb and flow of the tide can produce considerable currents in coastal estuaries and bays, reliable advance knowledge of the tides can be the difference between easy and hard exits from one's port. Reliable tidal prediction can make the difference between a successful return to one's home port or running aground in unexpected low water.

These factors make reliable long-term tidal prediction highly desirable. Since the behavior of the tides varies so much from place to place, the problem of prediction is best tackled as a curve fitting problem. Start with a good history of tides at each place on a coast. For each, find the curve that best fits the history and use it for prediction. Since the tides are periodic phenomena, one would expect that the families of functions to be fitted are based on the periodic trigonometric functions: sines and cosines of time. The natural method is straightforward Fourier analysis, which is the celebrated mathematical method for representing periodic functions in terms of sine and cosine harmonic constituents. To apply it, we would assume that the dependence of the water height on time t is given by the series:

$$a_0 + a_1\sin(t) + b_1\cos(t) + a_2\sin(2t) + b_2\cos(2t) + a_3\sin(3t) + b_3\cos(3t) + \ldots$$

with the series continued as far as needed. We know from the theory of Fourier series that any credible behavior of the tides over some nominated time can be represented arbitrarily closely by this expression. We merely need a suitable scaling for t, a suitable selection of the a and b coefficients and the inclusion of enough terms from the series.

While this is the obvious approach, in the past century and a half of work on tides, I have found no serious effort to provide this sort of analysis. The core difficulty, I conjecture, is that the dominant harmonic constituents present do not have the frequencies 1, 2, 3, … of the generic Fourier analysis. Combinations of these dominant constituents could still be captured by Fourier analysis with components of frequency 1, 2, 3, … However a large number of these components would be needed to represent accurately the summation of a few dominant harmonics whose frequencies are not in this set.

Instead, from the first moments, a physical basis has always been demanded for the harmonic constituents fitted to observed tidal histories. William Thomson (later Lord Kelvin) introduced the method of fitting harmonic constituents to tidal histories in 1867. Writing in his report to the British Association (Thomson, 1869), he noted that previous methods had merely recorded the times of high and low water. He proposed that fuller records be kept to which harmonic constituents would be fitted. The particular constituents he proposed were drawn directly from the background theory of the origin of the tides in the gravitational interaction of the earth, sun and moon.

The elements of this theory are widely known. The moon's gravity pulls on the waters of the earth's oceans. The pull is stronger than the average on the side of the earth nearer the moon and weaker on the side farther from the moon. The net effect is an elongation of the oceans into a spheroid that bulges away from the earth on both sides in line with the earth-moon axis, as shown in Figure 11.

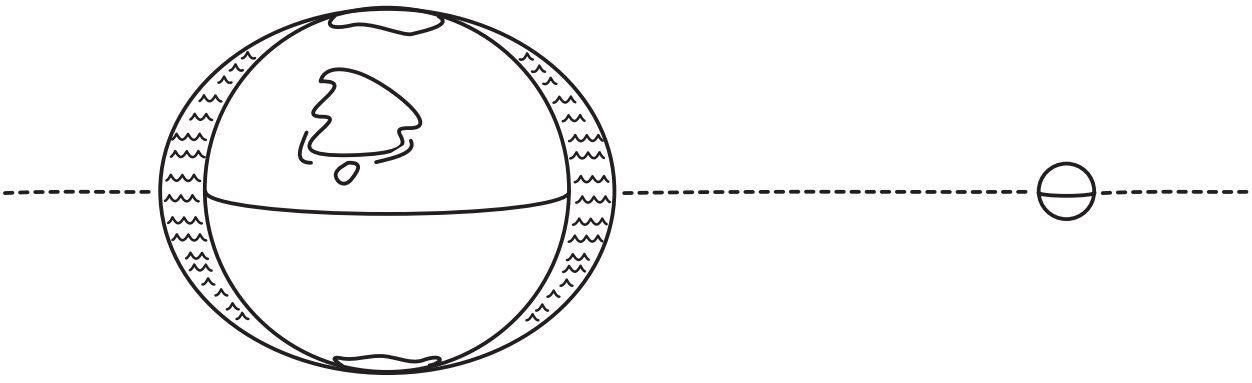


Figure 11. Tidal Bulge of Oceans Raised by the Moon.

The earth rotates daily under this bulged shape. As some location on the earth passes each bulge, that location registers a high tide. Since there are two bulges, each location registers two high

tides and two low tides each day. The cycle is only roughly daily since it completes when a point on the earth returns to its original position in relation to the moon. The moon orbits the earth once a month and moves in the same direction as the earth rotates. So to return to its starting position in relation to the moon, the earth must rotate slightly more than the full rotation of 24 hours. It requires roughly 24 hours and 50 minutes. In this time, two tide cycles are completed. Half this process gives us the most important harmonic constituent: the "principal lunar semi-diurnal [=half-daily]," written as M_2, where the 2 denotes two cycles per day. It has period of about 12 hours and 25 minutes.

Superimposed on this semi-diurnal cycle is another semi-diurnal cycle. It results from the gravitational attraction of the sun on the waters of the oceans. The sun's attraction also distorts the ocean waters into a spheroid elongated along the line of the earth-sun axis, or so it would if there were not greater distortions due to the moon's gravity. The bulge produced would be a little less than half that raised by the moon. It takes 24 hours exactly for a point on the earth to return to a position with same relation to the sun. There are once again two bulges passed in this time, so we cycle between them each 12 hours. This contributes another harmonic constituent, the "principal solar semi-diurnal" S_2, whose period is 12 hours.

That these two harmonic constituents have periods differing by about 25 minutes is of the greatest consequence for the tides. At the full or new moon, when the sun and moon align, the two bulges add and we have especially high tides, known as the "spring" tides. They are so named since more waters are imagined as springing forth. The effect is shown in Figure 12.

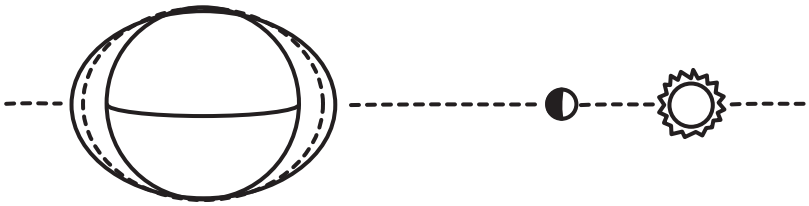


Figure 12. Spring Tides

Each 12 hours, the high water of lunar semi-diurnal cycle will lag behind that of the solar semi-diurnal cycle by about 25 minutes. This lag accumulates. After about a week, at the time of the half moon, the tidal bulges of the moon and sun are aligned roughly perpendicularly. The outcome is a lowering of the high tide and an elevation of the low tide, producing the more modest "neap" tides of Figure 13.

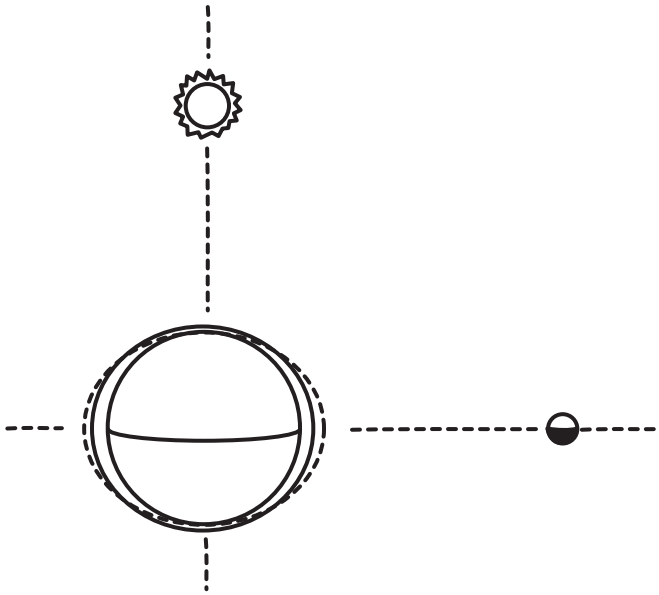


Figure 13. Neap tides

The combining of the two cycles to produce this further cycle of spring and neap tides is shown in Figure 14.

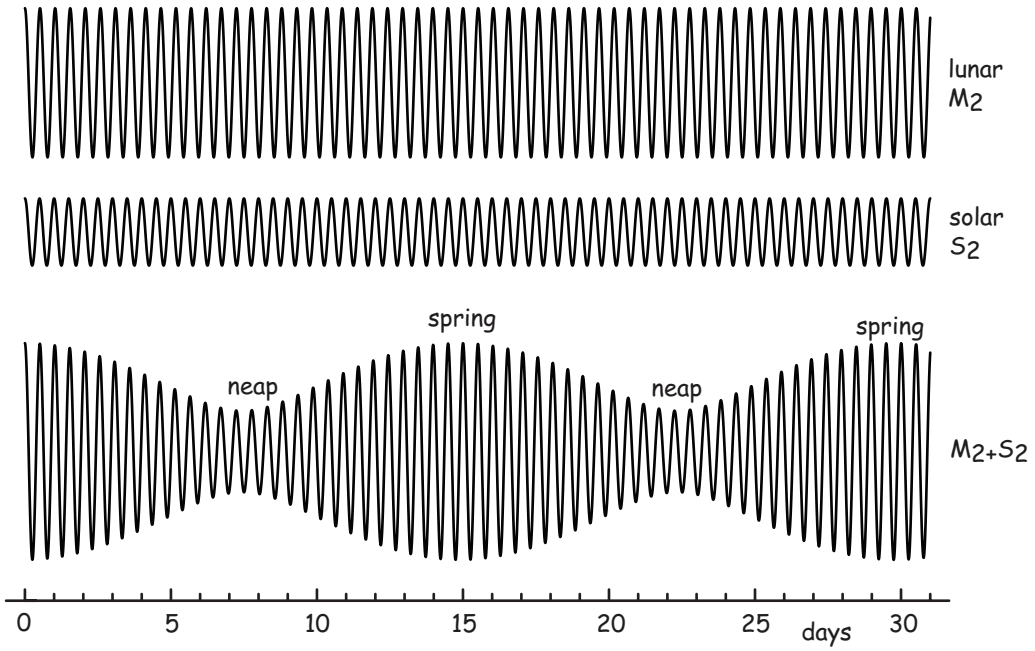


Figure 14. Spring and Neap Times from Combining Harmonic Constituents

This is one of the most important outcomes of the combining of the tidal harmonic constituents.

One might expect that there is little more to the harmonic analysis of the tides. What we have seen so far is adequate for an informal understanding of the gross behavior of the tides. It associates tidal levels with the phases of the moon in a qualitatively correct way. However it falls

far short of what is needed for successful quantitative prediction of the tides. Many more physical factors must be considered.

The sun and moon also move to the North and South, carrying their tidal bulges with them. In the course of a solar year, the sun completes one cycle around the ecliptic, moving between 23.5º North and 23.5º South of the stellar equator. The moon's monthly motion carries it along a plane that is inclined at about 5º to the plane of the ecliptic. Its resultant motions carry it North and South of the stellar equator, between maximum elongations from the stellar equator that vary from 18.5º to 28.5º. As the sun and moon change their longitude, they carry with them the tidal bulges that they raise. This affects the heights of the tides and does it differently at each location on the earth.

If the sun and moon are directly over the equator, the two tidal bulges will pass symmetrically over some fixed terrestrial location on the equator in the course of a day. In so far as these processes are concerned, successive tides will have equal height. If, however, the sun and moon have moved together to a position far to the North, then the two tidal bulges will be shifted towards different hemispheres. One will be massed in the Northern hemisphere and the other in the Southern hemisphere. As a result, a point on the earth away from the equator will meet with deeper and shallower portions of the successive bulges, adding a diurnal (daily) cycle to the semi-diurnal cycles so far mentioned. In extreme cases, the sun and moon pass overhead sufficiently far from the equator that, at some locations, one bulge may be missed entirely. These locations experience a single high tide per day. That is, their tides are on a diurnal cycle.

There are further complications. The size of the tidal bulge raised depends on the distance from the earth to both the sun and moon. Since the orbits are elliptical, the sun and moon approach and recede from the earth as they complete their cycles, annually and monthly, respectively. In addition, this cyclic effect is compounded by the perturbations induced by the sun on the moon's orbit. Those perturbations alter the eccentricity of the moon's ellipse introducing further variation in the distance of the moon from the earth. All these astronomical effects happen on regular cycles, readily predictable in advance. They are incorporated into tidal analysis by adding more harmonic constituents of the appropriate form.

These astronomical effects may seem overwhelming. However they are merely the most reliably regular of the influences on the tides. If the earth were a perfectly smooth spheroid, a tidal bulge of the ocean would wash over it as a uniform wave. However the earth is not a

perfectly smooth body and all sorts of irregularities in its surface obstruct the uniform passing of the tidal wave. These obstructions are great in coastal areas, which is precisely where we are seeking predictions. The problems are even worse if we wish to predict tides in bays and estuaries. For the rising and falling of the tide will be delayed by the need for water to flow in and out of the bay as the tidal wave passes. Enclosed bodies of water have their own natural frequencies with which water oscillates to and fro within them. The coming and going of tidal waves couples with these oscillatory process. All these processes are represented by further harmonic constituents.

The shallow-water constituents are of two types: overtides and compound tides. The first are the analog of harmonic overtones in music. For example, the principal lunar semi-diurnal M_2 consists of 2 high tides per day. It raises shallow-water overtides M_4 and M_6 with four and six peaks each day. Compound tides arise with a frequency that is the sum or difference of the components from which they are derived. The shallow-water terdiurnal MK_3 is derived from the principal lunar semi-diurnal M_2 and the lunar diurnal K_1. It sums their two and one peaks per day to give three peaks.

Finally, meteorological facts can have a major influence on tides. Strong winds can materially affect them. These factors, however, are the hardest to address. Accurate weather prediction is difficult even a day in advance, whereas tide tables are prepared years in advance. There is some small effort to allow for these meteorological effects by means of the solar components, S_a, S_{sa} and S_1; that is, the solar annual, solar semi-annual and solar diurnal, which have periods of a year, half a year and a day.[18]

In sum, the harmonic analysis of tides is complicated and difficult, even when we seek a sound physical basis for the harmonic constituents. Many are needed. This was already apparent to Thomson (1869, p. 491), who initially listed 23 constituents. Many more can be needed. The most difficult locations for prediction are complex estuaries, such as Anchorage, Alaska, and Philadelphia, Pennsylvania. An adequate analysis requires over 100 harmonic constituents. (Hicks, 2006 , p.40.) The United States National Oceanic and Atmospheric Administration

[18] For further discussion of these harmonic components, see Shureman (1958, pp. 39-48).

(NOAA) employs a standard set of 37 constituents for its tidal predictions for coastal regions in the US. Here is an illustration of their use.

The table shows the harmonic constituents used by NOAA for Annapolis, Maryland, in the Chesapeake Bay:[19]

	Constituent Symbol	Constituent Name	Amplitude	Phase	Speed
1	M2	Principal lunar semidiurnal	0.457	291.6	28.9841042
2	S2	Principal solar semidiurnal	0.071	319.5	30
3	N2	Larger lunar elliptic semidiurnal	0.095	270.5	28.4397295
4	K1	Lunar diurnal	0.194	356.7	15.0410686
5	M4	Shallow water overtides of principal lunar	0.012	58.3	57.9682084
6	O1	Lunar diurnal	0.157	6	13.9430356
7	M6	Shallow water overtides of principal lunar	0.011	159.6	86.9523127
8	MK3	Shallow water terdiurnal	0	0	44.0251729
9	S4	Shallow water overtides of principal solar	0	0	60
10	MN4	Shallow water quarter diurnal	0	0	57.4238337
11	NU2	Larger lunar evectional	0.021	268.5	28.5125831
12	S6	Shallow water overtides of principal solar	0	0	90
13	MU2	Variational	0	0	27.9682084
14	2N2	Lunar elliptical semidiurnal second-order	0.013	246.7	27.8953548
15	OO1	Lunar diurnal	0.006	347.3	16.1391017
16	LAM2	Smaller lunar evectional	0.011	318	29.4556253
17	S1	Solar diurnal	0.065	290.5	15
18	M1	Smaller lunar elliptic diurnal	0.011	1.2	14.4966939
19	J1	Smaller lunar elliptic diurnal	0.011	340.9	15.5854433
20	MM	Lunar monthly	0	0	0.5443747
21	SSA	Solar semiannual	0.119	44.5	0.0821373
22	SA	Solar annual	0.338	128.4	0.0410686
23	MSF	Lunisolar synodic fortnightly	0	0	1.0158958
24	MF	Lunisolar fortnightly	0	0	1.0980331

[19] Amplitude is measured in feet, phase in degrees and speed in degrees per hour. Source: http://tidesandcurrents.noaa.gov/data_menu.shtml?stn=8575512%20Annapolis,%20MD&type=Harmonic%20Constituents accessed August 10, 2012.

25	RHO	Larger lunar evectional diurnal	0.012	29	13.4715145
26	Q1	Larger lunar elliptic diurnal	0.025	331.6	13.3986609
27	T2	Larger solar elliptic	0.004	318.3	29.9589333
28	R2	Smaller solar elliptic	0.001	320.6	30.0410667
29	2Q1	Larger elliptic diurnal	0.004	15.1	12.8542862
30	P1	Solar diurnal	0.065	348.8	14.9589314
31	2SM2	Shallow water semidiurnal	0	0	31.0158958
32	M3	Lunar terdiurnal	0	0	43.4761563
33	L2	Smaller lunar elliptic semidiurnal	0.033	308.1	29.5284789
34	2MK3	Shallow water terdiurnal	0	0	42.9271398
35	K2	Lunisolar semidiurnal	0.021	317.9	30.0821373
36	M8	Shallow water eighth diurnal	0	0	115.9364166
37	MS4	Shallow water quarter diurnal	0	0	58.9841042

Table 1. Harmonic Constituents used by NOAA for Tidal Predictions at Annapolis, Maryland.

These thirty seven constituents fix the family of thirty seven component functions whose sum is to be fitted to the tidal history in Annapolis. Each consists of a cosine wave whose amplitude, phase and speed are to be determined either from background assumptions or by fitting to the tidal history. The resulting parameters, given in the last three columns of the table, are used to compute NOAA's tidal prediction. Figure 15 shows the result of combining them for the week of August 7, 2014.[20]

[20] The height predicted is above MLLW = mean lower low water. Source: http://tidesandcurrents.noaa.gov/noaatidepredictions/NOAATidesFacade.jsp?timeZone=2&dataUnits=1&datum=MLLW&timeUnits=2&interval=6&Threshold=greaterthanequal&thresholdvalue=&format=Submit&Stationid=8575512&&bmon=08&bday=07&byear=2014&edate=&timelength=weekly accessed August 10, 2012.

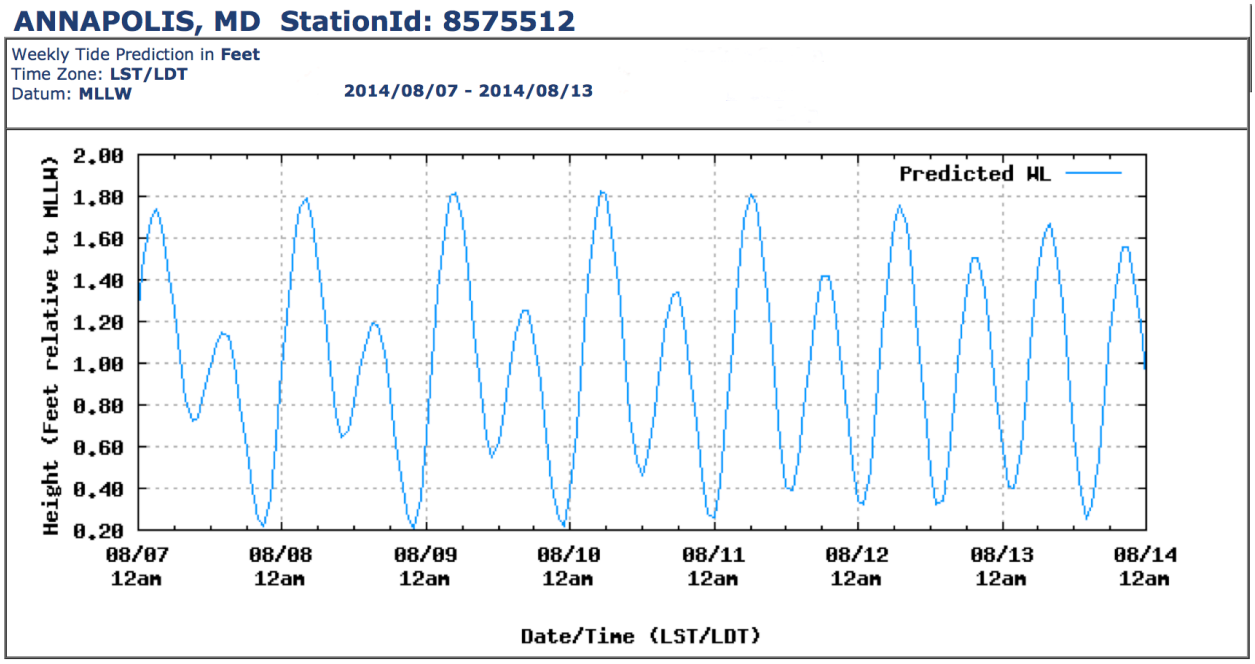


Figure 15. Tidal Prediction for Annapolis, August 7-13, 2014

References

Airy, George B. (1884) *Gravitation: An Elementary Explanation of the Principal Perturbations in the Solar System*. 2nd. Ed. London: MacMillan and Co.

Aquinas, Thomas (1945) *Basic Writings of Saint Thomas Aquinas*. Ed. Anton C. Pegis. Random House; Reprinted Indianapolis, IN: Hackett, 1997.

Brown, Ernest W. (1896) *An Introductory Treatise on the Lunar Theory*. Cambridge: Cambridge University Press.

Clark, Ronald W. (1984) *Einstein: the Life and Times*. New York: Avon.

Einstein, Albert (1915) "Erklärung der Perihelbewegung des Merkur aus der allgemeinen Realtivitätstheorie," *Königlich Preussische Akademie der Wissenschaften (Berlin), Sitzungsberichte* (1915), pp. 831-839.

Einstein, Albert (1933), "On the Methods of Theoretical Physics," in *Ideas and Opinions*. (New York: Bonanza, 1954), 270-276.

Einstein, Albert (1949), "Autobiographical Notes," in P. A. Schilpp, , ed., *Albert Einstein: Philosopher-Scientist*. Evanston, IL: Library of Living Philosophers.

Hicks, Steacy D. (2006) “Understanding Tides,” Center for Operational Oceanographic Products and Services, National Oceanic and Atmospheric Administration. http://www.co-ops.nos.noaa.gov/publications/Understanding_Tides_by_Steacy_finalFINAL11_30.pdf

Jeffreys, Harold (1916) “The Secular Perturbations of the Four Inner Planets,” *Monthly Notices of the Royal Astronomical Society*, **77**, pp.112-118.

Kelly, Kevin (2007) “A New Solution to the Puzzle of Simplicity,” *Philosophy of Science,* **74**, pp. 561-73.

Le Verrier, Urbain (1859), , "Lettre de M. Le Verrier à M. Faye sur la théorie de Mercure et sur le mouvement du périhélie de cette planète", *Comptes rendus hebdomadaires des séances de l'Académie des sciences (Paris)*, **49**, pp. 379–383.

Lorentz, Hendrik A. (1920) *The Einstein Theory of Relativity: A Concise Statement*. 3rd ed. New York: Brentano’s.

Mach, Ernst (1898) “The Economical Nature of Physical Inquiry,” pp. 186-213 in *Popular Scientific Lectures*. Trans. T. J. McCormack. 3rd ed. Chicago: Open Court.

Maurer, Armand A. (1999) *The Philosophy of Ockham in the Light of its Principles*. Toronto, Canada: Pontifical Institute of Mediaeval Studies.

Misner, Charles W.; Thorne, Kip S.; Wheeler, John Archibald (1973), *Gravitation*. San Francisco: W. H. Freeman

Moody, Lewis. F. (1944), "Friction factors for pipe flow", *Transactions of the American Society of Mechanical Engineers*, **66** (8), pp. 671–684.

Newcomb, Simon (1895) *The Elements of the Four Inner Planets and the Fundamental Constants of Astronomy*. Washington: Government Printing Office.

Newton, Isaac (1726), *Mathematical Principles of Natural Philosophy*. 3rd ed. Trans. Andrew Motte, rev. Florian Cajori. University of California Press, 1962.

Norton, John D. (2000), “ ‘Nature in the Realization of the Simplest Conceivable Mathematical Ideas’: Einstein and the Canon of Mathematical Simplicity,” *Studies in the History and Philosophy of Modern Physics*, 31, 135-170.

Popper, Karl (1968) *Logic of Scientific Discovery*. New York: Harper and Row.

Shureman, Paul (1958) *Manual of Harmonic Analysis of Tides*. Washington: United States Printing Office.

Sober, Elliott (1988), "The Philosophical Problem of Simplicity," CH. 2 in *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge, MA: Bradford, MIT Press.

Sterne, Theodore E. (1960) *An Introduction to Celestial Mechanics*. New York: Interscience Publishers.

Thomson, William (1869) "Committee for the purpose of promoting the extension, improvement, and harmonic analysis of Tidal Observations," *Report of the Thirty-Eighth Meeting of the British Association for the Advancement of Science, Norwich August 1868*. London: John Murray, Albemarle St. pp. 489-510.

Thorburn, W. M. (1918) "The Myth of Ockahm's Razor," *Mind*, **27**, pp. 345-53.

Watson, James C. (1861) *A Popular Treatise on Comets*. Philadelphia: James Challen & son.